OCCASIONAL PAPERS
THE MUSEUM
TEXAS TECH UNIVERSITY

NUMBER 122 18 OCTOBER 1988

MCZ LIBRARY
APR 13 1989
HARVARD UNIVERSITY

CONTIGUOUS CLUSTERING: A METHOD FOR IDENTIFICATION OF NONRANDOM AGGREGATES WITHIN POPULATION SAMPLES

RONALD K. CHESSER AND RONALD A. VAN DEN BUSSCHE

Investigations of the genetic substructure of natural populations have been a major concern for many years. Wright (1931) first proposed that organismal populations may be comprised of numerous subpopulations representing a mosaic of genetic identities. Small, isolated breeding groups within populations have important consequences regarding the rate and direction of evolutionary processes. Concomitant with the identification of population structuring, numerous statistical methodologies were developed. The *F*-statistics (Wright, 1951, 1969, 1978) were designed to quantify the relative amounts of genetic differentiation and nonrandom breeding of predefined populations. Subsequently, spatial autocorrelation analysis (Sokal and Oden, 1978*a*, 1978*b*) enabled *a posteriori* analysis of nonrandom distributions of genotypes within populations. Mantel (1967) and Sokal (1979) provided a technique of unbiased comparisons of matrices such as genetic distance versus geographic separation distance (see, Feder *et al.*, 1984). Mantel's method of matrix association testing has been extended to include multiple matrices (Smouse *et al.*, 1986). These techniques have made possible detailed description of population structure and importance of philopatry to evolution (Bengtsson, 1978; Greenwood, 1980; Chesser and Ryman, 1986). However, those methods do not provide a mechanism for the *a posteriori* identification of the individuals comprising nonrandom genetic aggregations. Selander (1970) attempted identifica-

tion of unique aggregates of genotypes of *Mus musculus* by isofrequency lines. Others have relied on political (for example, county and state lines, management units—Manlove *et al.*, 1976; Chesser *et al.*, 1980, 1982) or other artificial boundaries for the definition of meaningful biological groups.

Cliff and Ord (1981) and Ripley (1981) have presented numerous methods for examining geographic relationships of continuous and discretely distributed variables. Applications of these and other techniques have been used to define taxonomic boundaries (Lefkovitch, 1980, 1982, 1984) and population fission events (Smouse *et al.*, 1981); but, because they are based on phenotypic or genetic distance values, they have limited application to definition of subgroups within populations characterized by noncontinuous variables. Typically, character variation within populations is more conservative than that among populations and groupings based on phenetic distance may not accurately reveal underlying patterns.

The purpose of this paper is to present a method of identification of genotypic aggregates within populations in which the coordinates and genotypes of individuals are known. As this method forms potential clusters only for individuals that are spatially contiguous, we refer to the method as "contiguous clustering." We have analyzed, as an example, the genotypic substructure of a population of the black-tailed prairie dog (*Cynomys ludovicianus*).

Materials and Methods

Data for genetic markers and exact position of individual prairie dogs (*Cynomys ludovicianus*, N = 77) within a large population near Portales, New Mexico, were available from a previous study (see Chesser, 1983, PORT population). Three loci were found to have sufficient variability for further statistical analyses: Purine nucleoside phosphorylase (PNP, Enzyme Commission number (E.C.) 2.4.21); Phosphoglucomutase-2 (PGM-2, E.C. 5.4.2.2); and Phosphogluconate dehydrogenase (PGDH, E.C. 1.1.1.44). The exact collection location of each prairie dog was recorded on a map of the population, which represented 560 by 500 meters divided into 8.25 meter grids. The linear distance between each pair of individuals subsequently was calculated. Boundaries of each breeding group (coterie) were previously estimated (Chesser, 1983) and each individual was noted as to its coterie of origin (if any).

Contiguous Clustering Method

Below we describe a method of identifying clusters for each genotype within populations in which the x-y coordinates for each individual are known. The method is designed to build subgroups of individuals that are spatially contiguous. No new statistical methods are derived, but rather a sequence of methodologies are provided, using previously described statistics, to define the contiguity of individuals possessing discrete character values.

The hypergeometric probability distribution long has been recognized as a valuable statistic for describing the frequency of specific characteristics from a subsample of known (or at least estimated) population constituency, particularly when the size of the subsample relative to the total sample is not small (Hahn and Shapiro, 1968:151-152). The parameters necessary for the calculation of the hypergeometric probability are: (1) the size of the total population sampled (N); (2) the number of individuals within the population possessing the designated discrete character to be tested (A—in this instance the characters are specific genotypes); (3) the number of individuals in the population possessing discrete characters other than the designated character (B—"other" genotypes, $B = N - A$); (4) the number of individuals within a subset (potential cluster) of the total population (n); (5) the number of individuals within the subset that possess the designated character (a); and (6) the number of individuals within the subset that possess alternative characters ($b = n - a$). The probability that a subsample of size n includes a individuals possessing the designated character is estimated as

$$\Pr(a,b \mid A,B) = \frac{\binom{A}{a}\binom{B}{b}}{\binom{N}{n}}$$

In many instances, the binomial distribution will be more suitable for assessment of aggregate significance than the hypergeometric because 1) aggregates within population samples may be small relative to the total sample size, and 2) statistical significance of the aggregates may be readily obtained. Using the same notation as that above, if we assume that the population represents a random sample, the probability of a particular genotype is $p = A/N$, and the observed proportion in the subsample is a/n. The cross entropy is

$$H = \frac{a}{n} \log_e \left[\frac{a}{np}\right] + \frac{b}{n} \log_e \left[\frac{b}{n(1-p)}\right]$$

and 2nH has an asymptotic X^2 distribution with one degree of freedom. This statistic measures the distance from an assumed random distribution with parameters determined by A and B.

We utilize the hypergeometric distribution and cross entropy X^2 to test for clusters of individuals possessing specific genotypes. Clusters are defined as subgroups that contain significantly greater numbers of a given character than from a random subset of the population. It is assumed that all individuals in the population have equal probabilities of being included in the sample under investigation. In some circumstances, this assumption may not be met, such as when individuals possessing different genotypes differ in their behavior or when investigators confine their samples to easily accessible regions (for example, roadsides) rather than a random sample of occupied area.

The criteria of subset formation are similar to those of conditional clustering (Lefkovitch, 1982) in that 1) if the maximum distance between individuals in a subset approaches zero, then the number of individuals in the subset approaches unity, and 2) if the maximum distance between individuals in a subset approaches infinity, then all individuals under study belong in the subset. In this paper, we satisfy these conditions by considering individuals to be of the same subset (that is, are contiguous) if they fall within an ellipse with the foci (f1 and f2) being the positions of the two most distant individuals within a subset (subset formation is described below). To equalize the contribution of the foci individuals to the constituency of the subgroup, we define the eccentricity of the ellipse as inversely proportional to the ratio of the distance ($d_{f1,f2}$) separating the two foci and the maximum distance (d_{max}) separating any pair of individuals in the population. Therefore, if $d_{f1,f2}/d_{max} = 1$, then the result will be an ellipse that encloses all individuals in the population sample; as $d_{f1,f2}/d_{max}$ approaches zero, the ellipse will flatten to include fewer surrounding samples. The eccentricity of the ellipses need not be calculated directly to determine the samples that fall within or on the ellipse. The length of the major axis (M) that passes through the two foci can be calculated as $M = 2[(d_{f1,f2}/2)^2 + (d_{f1,f2}{}^2/2d_{max})^2]^{1/2}$. For any other sample (p) to be included within the ellipse the condition, $d_{f1,p} + d_{f2,p} \leq M$ must be satisfied.

The necessary steps for identifying and testing subsets of contiguous individuals within the population follow. First,

calculate the maximum distance (d_{max}) separating any pair of individuals; (2) locate a single individual that possesses the designated genotype. The position of this individual is the designated reference point; (3) locate the individual nearest to the reference point that possesses the same designated genotype. This individual will be a focus (f1) of the ellipse and these two individuals form the initial subset; (4) find the member of the subset most distant from f1 (initially this is the designated reference point). The position of this individual is the second focus (f2) of the ellipse; (5) calculate the length of the major axis (M) through the two foci; (6) include, as members of the subset, all individuals, regardless of genotype, falling within or on the ellipse around the major axis; (7) calculate the number of designated genotypes (a) and "other" genotypes (b) within the subset (size n); (8) calculate the probability of forming the above subset using the hypergeometric distribution, or alternatively, calculate the cross entropy X^2 value. However, because we have necessarily formed subsets containing two individuals possessing the genotype in question, the values of A, a, N, and n must be replaced by A-2, a-2, N-2, and n-2, respectively. If the probability is less than a defined alpha level, then the members of the subset are designated as comprising a cluster—otherwise no cluster is defined by this subset; (9) find the next nearest individual possessing the designated genotype from the reference point (distance from the reference point is greater than that found previously in step three); (10) steps four through nine are repeated until all individuals in the total sample possessing the designated genotype have been included in the subset. It is important to note that the size of the subset continues to increase, including individuals from all previous subsets; however, only those subsets that had probabilities less than the desired alpha level are retained as clusters; (11) use another individual possessing the designated genotype as the reference point; (12) steps three through 11 are repeated until all individuals possessing the designated genotype have been used as the reference point. Although identical subsets may be tested repetitively by steps 11 and 12, this procedure insures that all possible combinations of shapes of subsets for a particular genotypic distribution have been tested for cluster conformity; (13) repeat steps two through 12 for each genotype of each locus.

The methods described above allow contiguity, and thus the shapes of the ellipses, to be defined by the data. Foci that are relatively far apart will determine more circular ellipses, and

therefore encompass a larger number of surrounding individuals, than will foci that are close together. Therefore, the criteria for contiguity will differ among genotypes and among loci. In reality, most organisms do not occupy a single coordinate, but rather a range of spatial positions. When data for such activity areas for individuals are available, the matrix of separation distances between individuals may be modified as proposed by Lefkovitch (1984:494).

Independently segregating loci may not depict identical population topographies even in highly structured populations (Michod and Anderson, 1978; Jacquard, 1974). Therefore, the overall structure of the population may be represented best by the combination or overlaying of clusters found for the various genotypes and loci. The alpha level can be designated according to the wishes of the investigators. In the following analyses of contiguous clusters in the population of *C. ludovicianus*, we identified clusters of genotypes with alpha levels less than, or equal to, 0.05 for the cross entropy X^2 and probabilities ≤ 0.005 for hypergeometric values (hereafter, we refer to the statistical probability associated with the X^2 and the probability for the hypergeometric as *P* and *Pr*, respectively). We also tested for clusters of individual alleles within the population at the same values.

Results

A total of nine contiguous clusters at $Pr \leq 0.005$ were found within the *C. ludovicianus* population (Fig. 1). Four of the clusters were detected for the PNP locus (two for the FF genotype, and one each for the SF and MF genotypes). For the PGDH locus, four contiguous clusters were detected with two resultant from the high incidence of the SF genotype and one each for the FF and SS genotypes. The contiguous clustering method was able to detect a single aggregate of MS genotypes for the PGM-2 locus. A total of 32 (41.6 percent) individuals was included in more than one cluster. However, none of the clusters found for a locus overlapped and no identical clusters were found for the separate loci. The cluster formed by the PNP-MF genotypes included only two additional individuals not found in the PGDH-SS genotype cluster. Two clusters, one for PNP-FF (upper left) and PGM-MS, were circumscribed by the large PGDH-FF cluster. Only nine (11.7 percent) prairie dogs were not included in any cluster. Five of the clusters included individuals

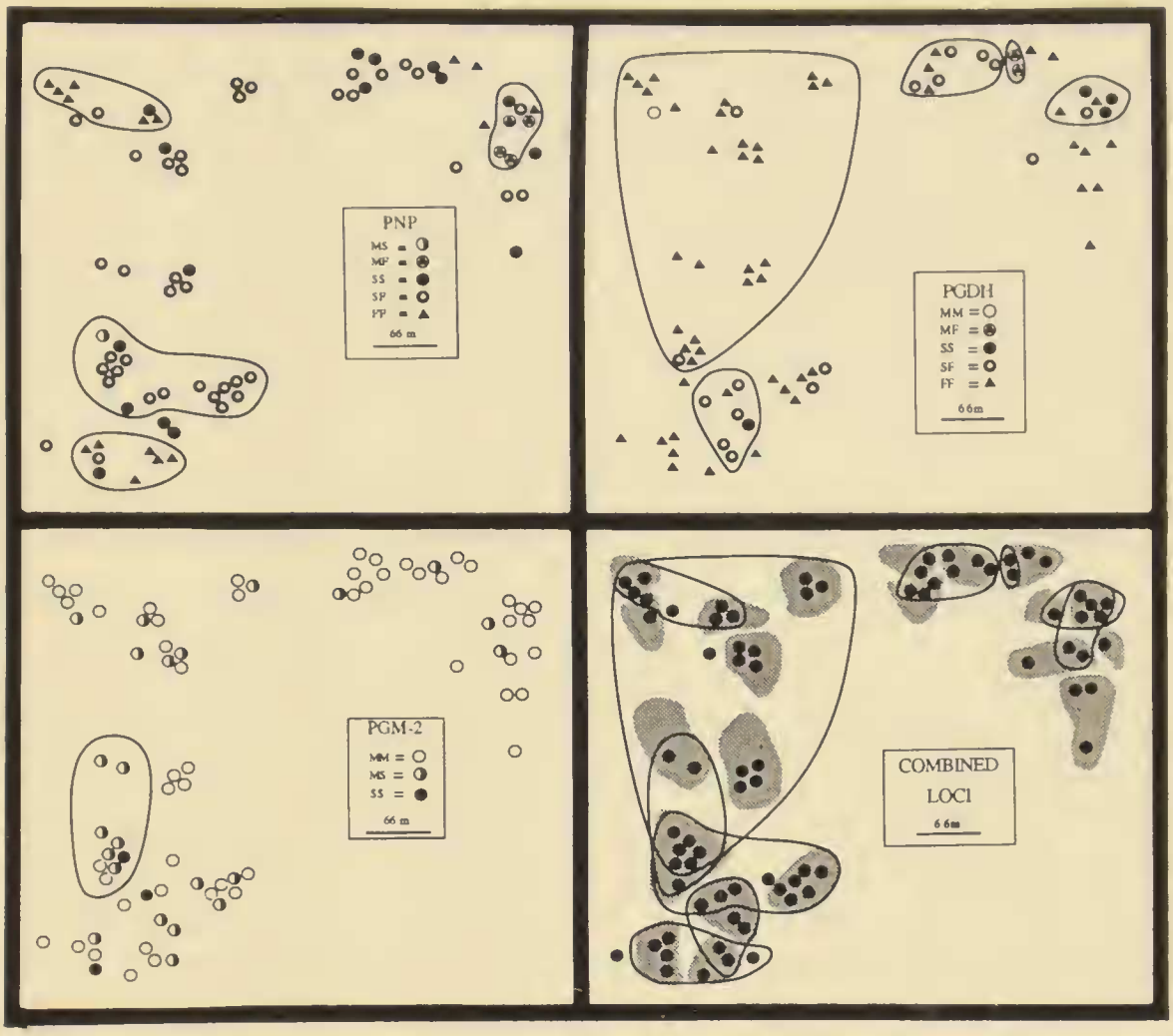

FIG. 1.—Graphical representation of the position of individuals possessing specific genotypes within a population of *Cynomys ludovicianus*, and the results of contiguous clustering analysis for each locus (see text for locus abbreviations) and for all loci combined (cross hatched area delineates coterie boundaries). The solid lines represent contiguous clusters of individuals from the hypergeometric distribution ($Pr \leq 0.005$). All clusters for the PNP locus and the large cluster for PGDH-FF genotypes were also significant when analyzed by the cross entropy X^2 method ($P \leq 0.05$).

from two coteries, three clusters traversed three coteries, and a single cluster contained individuals from eight separate coteries (coteries are identified by the hatched areas on the combined loci—Fig. 1). The average cluster length was 107.5 ± 29.7 meters.

Only five of the clusters identified by the hypergeometric distribution were found to be significant ($P < 0.05$) when analyzed by the cross entropy X^2 from the binomial. These clusters included all four found for the PNP locus and the large cluster formed by the PGDH-FF genotype. Twenty-two (28.6 percent) prairie dogs were not included in any significant cluster. No significant clusters were identified when the analysis was performed on the distribution of individual alleles for the three loci.

Discussion

It has been documented previously (Chesser, 1983) that considerable segregation of genotypes exists in *C. ludovicianus* populations. The contiguous clustering method provides the advantage of identifying the nonrandom assemblages of genotypes or alleles, or both, within the population. The potential vagility of prairie dogs is sufficient to enable an individual easily to traverse the sampling area in a short period of time. It appears from the clustering information that behavioral or historical factors, or both, have not resulted in adequate genetic exchange to produce a spatially random assemblage of genotypes throughout the population.

The results of the contiguous clustering were strikingly different for genotypic and allelic data. Individual alleles for the three loci are evenly distributed throughout the population, whereas genotypes are nonrandomly associated into local aggregates. This result would appear to indicate that no segment of the population has been isolated from others for extended periods of time, thereby leading to the segregation of unique genic characters; rather, it appears that the temporary maintenance of separate breeding groups, together with low effective sizes of local segregates, may produce discrete groups of particular genotypes. Because the coteries of black-tailed prairie dogs usually are comprised of a single breeding male and four to six breeding females (King, 1955), the identity or difference in the genotypes of the male and females will have a dramatic impact on the local genotypes of the next generation. That is, if the male was a homozygote for a different allele than that of the majority of females, then most of the offspring would have a heterozygous genotype. Conversely, a high proportion of homozygous individuals could be produced in a similar fashion. This pattern of localized genotypic distribution would be emphasized by the strong philopatry of females (Hoogland, 1985) to their native coterie.

The differences of aggregate identification for the hypergeometric and binomial methods demonstrate the difficulties in the interpretation of statistical significance from the probabilities derived from the hypergeometric distribution. Four of the clusters identified with the hypergeometric distribution for $Pr < 0.005$ were not statistically significant ($P < 0.05$) when analyzed by the cross entropy X^2. The X^2 values are measures of deviation from random distribution of characters within subsets relative to their

distributions in the total population, and hence, are more readily interpreted in a statistical context than are probabilities associated with the hypergeometric method. When subsample size (n) is large relative to the total sample size (N), the probabilities associated with the hypergeometric may be more accurate than those of the binomial (Hahn and Shapiro, 1968:152), although the differences are usually small.

The average distances between individuals within the contiguous clusters for genotypes (Fig. 1) were similar to, but consistently greater than, the significant span distances of nominal spatial autocorrelation (Sokal and Oden, 1978*a*, 1978*b*) analyses of these data (Fig. 2). However, these values need not necessarily coincide as the two methods determine "joins" in a different fashion and are based on different distributional properties of the data. Also contiguous clustering, as described herein, does not consider aggregations of unlike (heterogeneous) genotypes. Few significant clusters would be likely for populations comprised predominantly of heterogeneous subgroups, and the two methods thus may show disparate results. Contiguous clustering however, provides the advantage of identification of small aggregates for which spatial autocorrelation may lack sufficient statistical power for resolution.

The manner in which the animals are collected and the coordinates assigned may affect the results of clustering analyses. In this study, we have assumed that the points of capture are indicative of the normal area occupied by an individual. Deviations from that assumption could lead to either dilution or augmentation of clusters. We feel that such biases usually would act to disintegrate existing clusters rather than create artificial clusters. However, for the prairie dogs, we are not confined by individual capture points because the coterie boundaries also represent the normal activity areas. If we assign the coordinates of each member of a coterie as those of the center of the coterie (see, Lefkovitch, 1984:494), then the contiguous clustering may be performed on the basis of activity regions. This is a somewhat unusual analysis, due to the coincidental centers of activity for all coterie members, as it actually represents a contiguous clustering of the coteries themselves. The results of the clustering procedure (Fig. 3) indicates a large aggregation of 16 coteries with significantly greater PNP-SF genotypes than expected, two smaller clusters containing two coteries (also for PNP genotypes), and one cluster comprised by a single coterie containing a

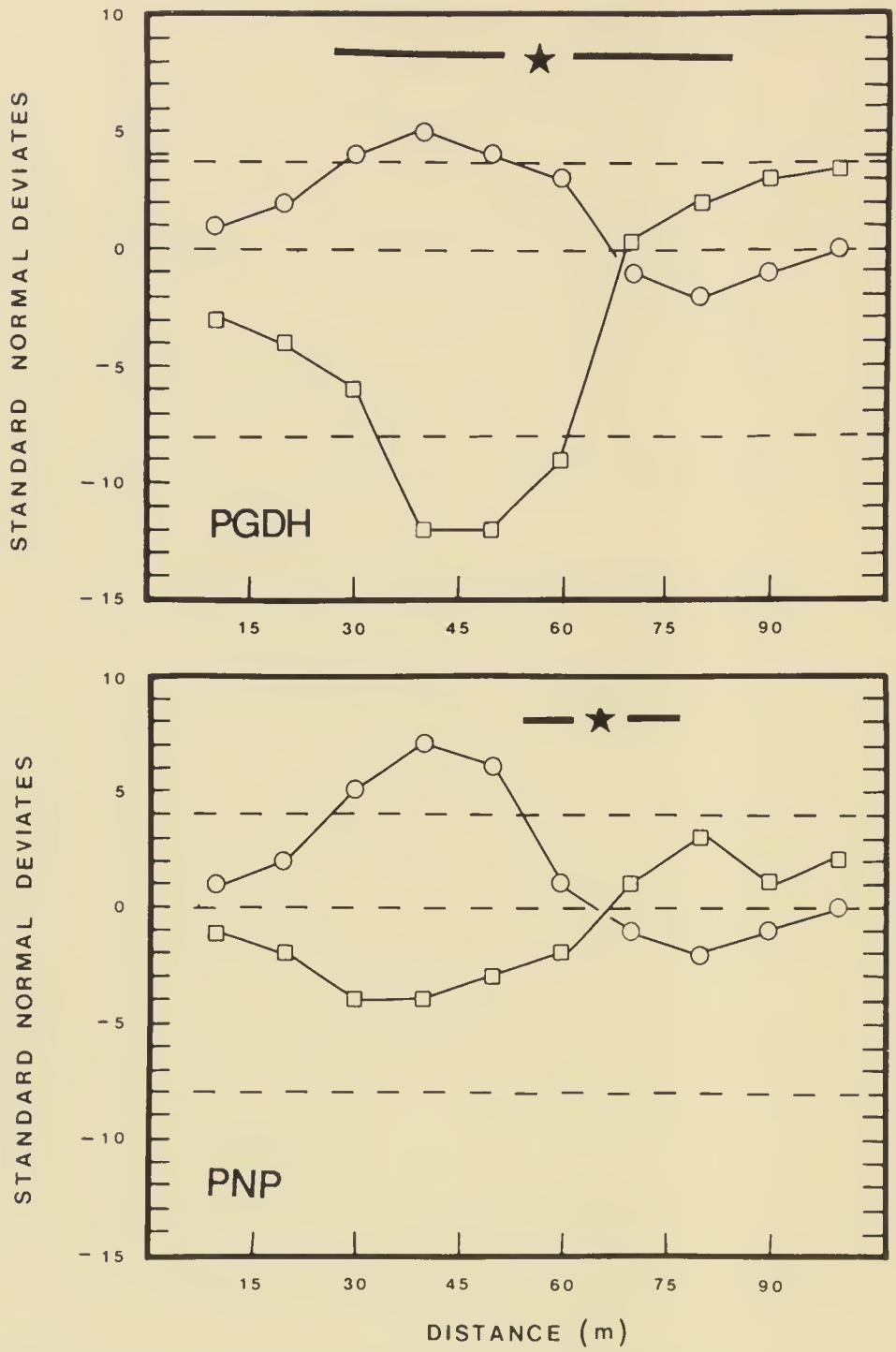

FIG. 2.—Correlograms representing the results of nominal spatial autocorrelation analyses of the genotypes of nucleoside phosphorylase (PNP) and 6-phosphogluconate dehydrogenase (PGDH) for a population of *Cynomys ludovicianus*. The open circles represent values for "joins" of like genotypes, whereas the open blocks represent values for unlike "joins." The upper and lower dashed lines represent the critical limits of the standard normal deviates beyond which values are considered significant ($P \leq 0.05$). Average distance separating individuals within clusters and the standard error is represented by the star and line at the top of the figure.

significant group of PGDH-SS genotypes. Therefore, although there appears to be sufficient gene flow within the population to prevent the segregation of individual alleles, the significant

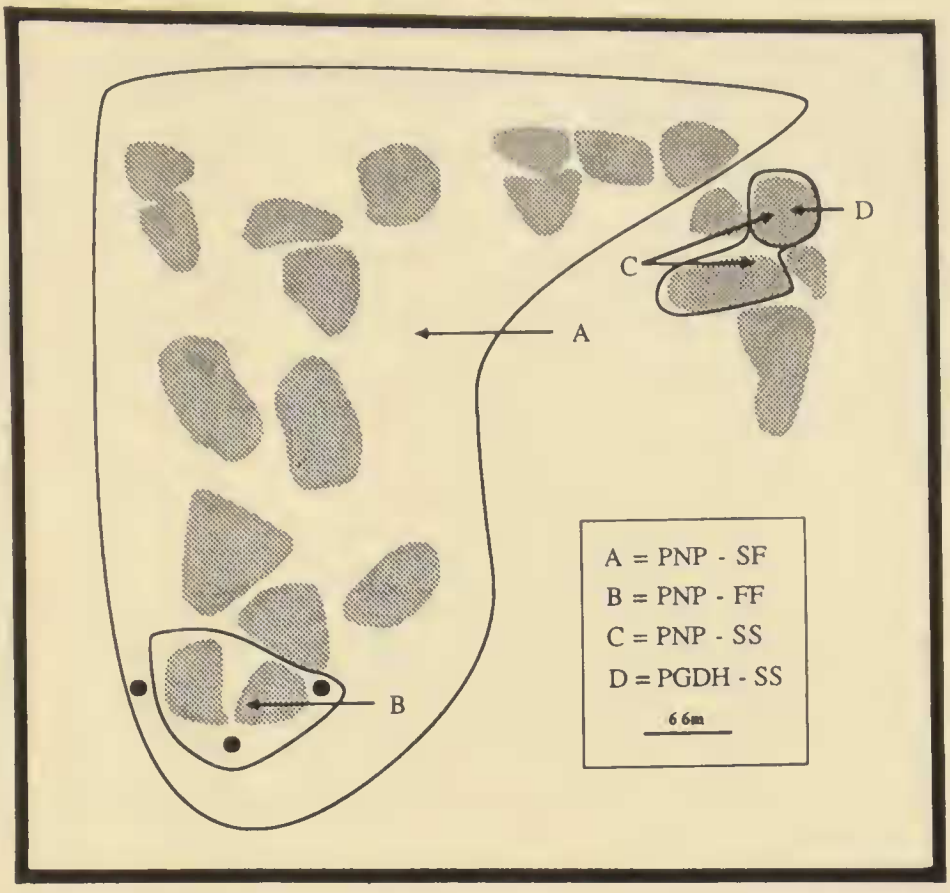

FIG. 3.—Graphical representation of the center of activity (coteries) of individuals possessing specific genotypes within a population of *Cynomys ludovicianus*, and the results of contiguous clustering analysis for each locus (see text for locus abbreviation). The solid lines represent clusters of coteries at $Pr \leq 0.005$. The black dots represent individuals not assigned to a coterie.

genotypic association of contiguous coteries indicates the probability of nonrandom genetic exchange among coteries.

It should be emphasized that the clusters identified by the contiguous clustering method do not necessarily represent segregated breeding groups. Lack of independence of genotypic distributions may be achieved in a variety of ways. Extended family lines represented by the offspring of a single parent or highly related parents may produce spurious, or at least temporary, clusters of genotypes. Subsequent dispersal of the offspring may eliminate the observed genotypic aggregation. Obviously, selection acting on a geographic, or microgeographic, scale can affect the pattern of observed character distributions. As with any analytical technique, interpretation of the significance of contiguous clusters should be made only with the consideration of the appropriate methodological caveats and the biology of the organism studied.

Contiguous clustering provides a valuable method of identifying the position of unique clusters of individuals within

populations. This method should have applications in various disciplines of biological and physical sciences because data need not be genotypes but may be of any noncontinuous form, and individuals may be represented simply by map coordinates. Also, the matrices of distances separating individuals may be transformed to reflect more parsimonious dispersal routes rather than linear distance (for example, river systems—Sokal and Oden, 1978*a*). Future applications of the contiguous clustering method undoubtedly will explore further conventions for the identification of population structure.

Acknowledgments

We wish to thank R. J. Baker, R. D. Bradley, M. J. Hamilton, C. S. Hood, D. A. McCullough, R. D. Owen, J. Sites, and C. A. Porter for their critical reviews of previous versions of this manuscript. Special thanks are extended to Dr. L. P. Lefkovitch for his many suggestions and methodological commentary. Funding for this study was provided in part by the New Mexico Department of Game and Fish and a grant from the National Science Foundation (BSR-86-00646) to R. J. Baker and Chesser.

Literature Cited

Bengtsson, B. O. 1978. Avoid inbreeding: at what cost? J. Theor. Biol., 73:439-444.

Chesser, R. K. 1983. Genetic variation within and among populations of the black-tailed prairie dog. Evolution, 37:320-331.

Chesser, R. K., and N. Ryman. 1986. Inbreeding as a strategy in subdivided populations. Evolution, 40:616-624.

Chesser, R. K., M. H. Smith, and I. L. Brisbin, Jr. 1980. Management and maintenance of genetic variability in endangered species. Internat. Zoo Yearbook, 20:147-154.

Chesser, R. K., M. H. Smith, P. E. Johns, M. N. Manlove, D. O. Straney, and R. Baccus. 1982. Spatial, temporal and age-dependent heterozygosity of beta-hemoglobin in white-tailed deer. J. Wildlife Manag., 46:983-990.

Cliff, A. D., and J. K. Ord. 1981. Spatial processes. Models and applications. Pion Limited, London, 266 pp.

Feder, J. L., R. K. Chesser, M. H. Smith, W. J. W. Godt, and K. Asbury. 1984. Biochemical genetics of mosquitofish. II. Demographic differentiation of populations in a thermally altered reservoir. Copeia, 1984:108-119.

Greenwood, P. J. 1980. Mating systems, philopatry and dispersal in birds and mammals. Anim. Behav., 28:1140-1162.

Hahn, G. J., and S. S. Shapiro. 1968. Statistical models in engineering. John Wiley and Sons. New York, 335 pp.

Hoogland, J. L. 1985. Infanticide in prairie dogs: lactating females kill offspring of close kin. Science, 230:1037-1040.

Jacquard, A. 1974. The genetic structure of populations. Springer-Verlag, New York, 569 pp.

King, J. A. 1955. Social behavior, social organization, and population dynamics in a black-tailed prairie dog town in the Black Hills of South Dakota. Contrib. Lab. Vert. Biol., Univ. Michigan, 67:1-123.

Lefkovitch, L. P. 1980. Conditional clustering. Biometrics, 36:43-48.

———. 1982. Conditional clusters, musters, and probability. Math. Biosci., 60:207-234.

———. 1984. A nonparametric method for comparing dissimilarity matrices, a general mesure of biogeographical distance, and their application. Amer. Nat., 123:484-499.

Manlove, M. N., M. H. Smith, H. O. Hillestad, S. E. Fuller, P. E. Johns, and D. O. Straney. 1976. Genetic subdivision in a herd of white-tailed deer as demonstrated by spatial shifts in gene frequencies. Proc. Ann. Conf. Southeast. Assoc. Game Fish Comm., 30:407-492.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res., 27:209-220.

Michod, R. E., and W. W. Anderson. 1979. Measures of genetic relationship and the concept of inclusive fitness. Amer. Nat., 114:637-647.

Ripley, B. D. 1981. Spatial statistics. John Wiley and Sons, New York, 252 pp.

Selander, R. K. 1970. Behavior and genetic variation in natural populations. Amer. Zool., 10:53-66.

Smouse, P. E., J. C. Long, and R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool., 35:627-632.

Smouse, P. E., V. J. Vitzthum, and J. V. Neel. 1981. The impact of random and lineal fission on the genetic diveregence of small human groups: a case study among the Yanomama. Genetica, 98:179-197.

Sokal, R. R. 1979. Testing statistical significance of geographic variation patterns. Syst. Zool., 28:227-232.

Sokal, R. R., and N. L. Oden. 1978*a*. Spatial autocorrelation in biology. 1. Methodology. Biol. J. Linn. Soc., 10:199-228.

———. 1978*b*. Spatial autocorrelation in biology. 2. Some biological implications and four applications of evolutionary and ecological interests. Biol. J. Linn. Soc., 10:229-249.

Wright, S. 1931. Evolution in Mendelian populations. Genetics, 16:97-159.

———. 1951. The genetical structure of populations. Ann. Eugenics, 15:323-354.

———. 1969. Evolution and genetics of populations. Vol. 2. The theory of gene frequencies. Univ. Chicago Press, Chicago, 511 pp.

———. 1978. Evolution and the genetics of populations. Vol. 4. Variability within and among natural populations. Univ. Chicago Press, Chicago, 580 pp.

Address of authors: *Department of Biological Sciences and The Museum, Texas Tech University, Lubbock, Texas 79409.* *Received 8 February, accepted 4 April 1988.*